the **ca**ncer **B**iomedical **I**nformatics **G**rid (caBIG):

Building the World Wide Web for Cancer Research

Ken Buetow

NCICB/NCI/NIH/DHHS

NCI biomedical informatics

- ***Goal:*** A virtual web of interconnected data, individuals, and organizations redefines how research is conducted, care is provided, and patients/participants interact with the biomedical research enterprise

Cancer Biomedical Informatics Grid (caBIG)

- Common, widely distributed infrastructure permits cancer research community to focus on innovation
- Shared vocabulary, data elements, data models facilitate information exchange
- Collection of interoperable applications developed to common standard
- Raw published cancer research data is available for mining and integration

NATIONAL INSTITUTES OF HEALTH
NATIONAL CANCER INSTITUTE

caBIG pilot status - organizational

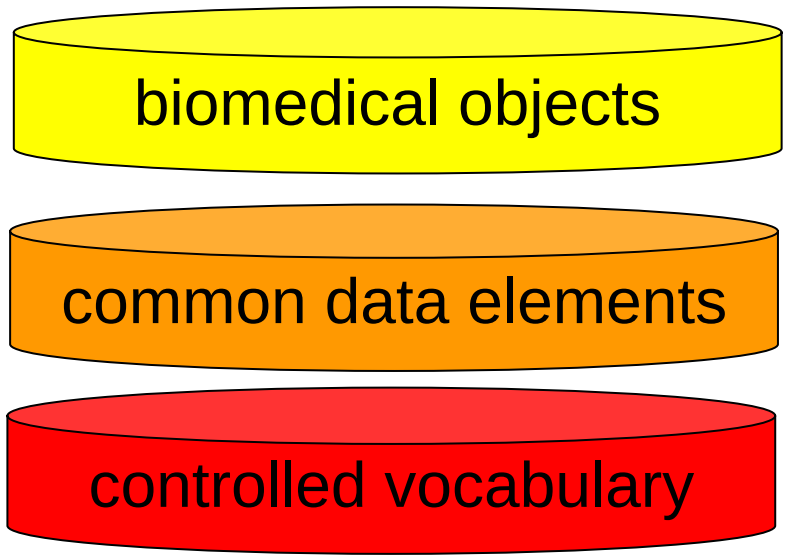
- Workspaces
  - Clinical Trials Management System
  - Integrated Cancer Research
  - Tissue Banks and Pathology
  - Vocabulary and Common Data Elements
  - Architecture
- Strategic Working Groups
  - Data Sharing and Intellectual Capital
  - Training
  - caBIG Strategic Planning
- Special Interest Groups
  - 23 groups focused on specific topics

caBIG pilot status - participation

- **Pilot** – NCI designated Cancer Centers
  - Members: 44 institutions – executed base agreements
    - developers
    - adopters
    - working group members
    - Statistics
      - » Over 450 active participants
      - » 196 teleconferences
      - » 10 face-to-face meetings
  - Volunteers
    - academic Centers
    - industry
- **Partners**
- **Affiliates**

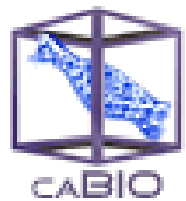
cancer **C**ommon **O**ntologic **R**epresentation **E**nvironment (**caCORE**)

- Information integration
- Cross-discipline reasoning



Standards supporting infrastructure

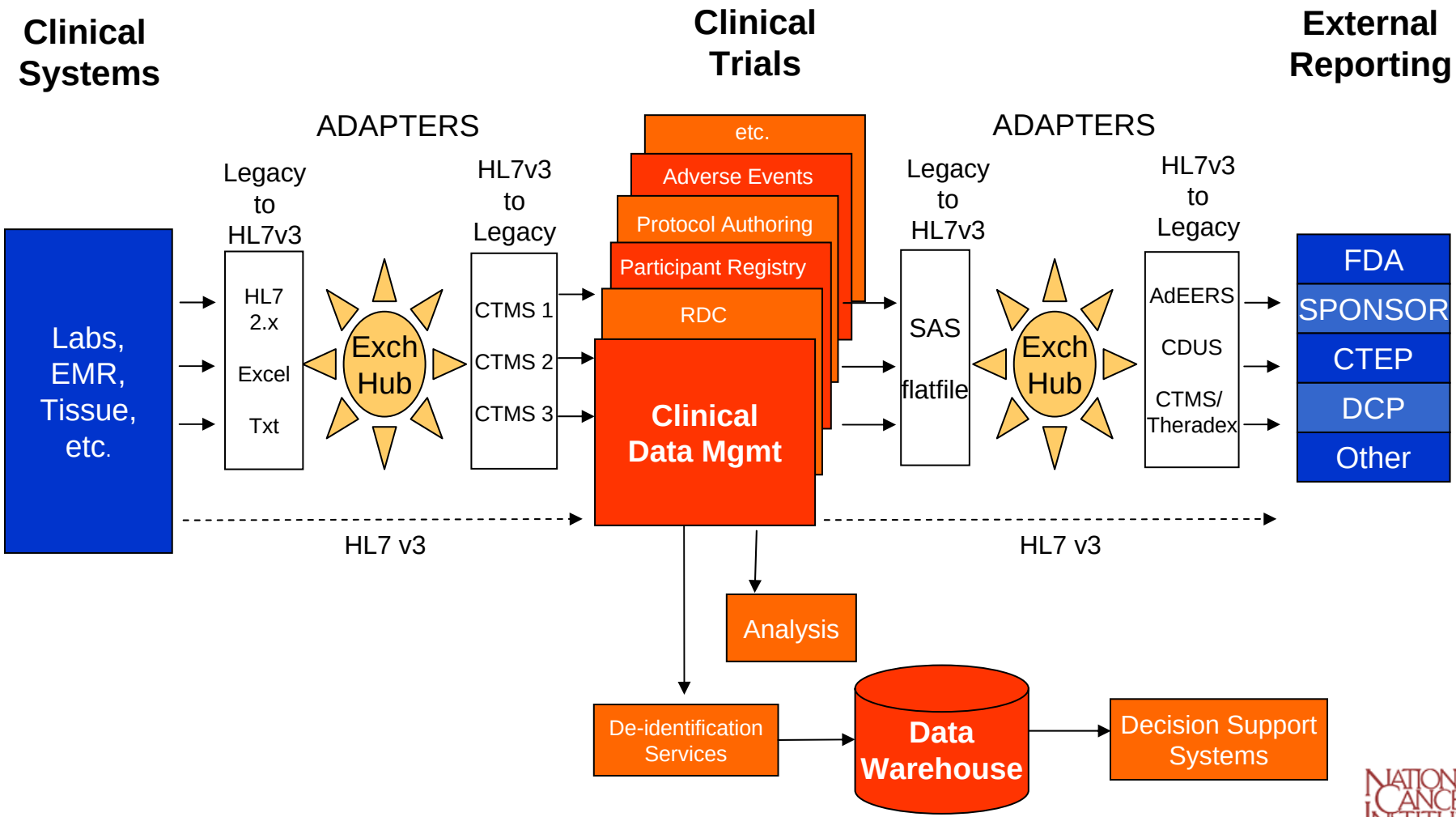
- ***E**nterprise **V**ocabulary **S**ervices (**EVS**)*
  - *Browsers*
  - *APIs*
- ***ca**ncer **B**ioinformatics **I**nfrastructure **O**bjects (**caBIO**)*
  - *Applications*
  - *APIs*
- ***ca**ncer **D**ata **S**tandards **R**epository (**caDSR**)*
  - *CDEs*
  - *Case Report Forms*
  - *Object models*
  - *ISO 11179 model*



caBIG Compatibility Matrix

Maturity Model	LEGACY	BRONZE	SILVER	GOLD
Interface Integration	*-No Programming interfaces to the system are available. Only local data files in a custom format can be read* ***-- Some ad hoc data transfer mechanism such as FTP***	***- Provide baseline* programmatic access to data. Data can be read from remote electronic sources or from commonly used file formats Data can be pushed out to from applications to other external data sources.***	*-- Well-described API's that provide access to data objects.* *-- System architecture separated into tiers and interoperable components* *-- Data read in from standards-based electronic sources that support standard or commonly used interchange formats* *-- Documented component description of the underlying data structures that are accessible* ***-- Standard messaging systems where appropriate***	*- All features of Silver, plus:* *- Interoperable with data grid architecture to be defined by caBIG* ***- Fully componentized provide access to individual resources in the form of grid services***
Vocabularies / Terminologies & Ontologies	***- Free text used throughout for data collection***	***- Use of publicly accessible standardized controlled vocabularies as well as local terminologies***	***- Messages need to be based on a common messaging model (such as HL7)***	*- All features of Silver, plus:* *- Fully compliant with caBIG recommended standards for vocabulary terminology services and content sources* ***- Common Data Elements should be built using controlled terminology***
Data Elements	***- No Structured metadata is recorded***	***- Some type of metadata describing the information in the system is used for data collection and external reporting. Metadata is retrieved from external repository shared by multiple applications.***	*- Use common standard electronic representation for CDE's such as ISO 11179 or comparable standard* ***- CDEs are harmonized and re-used from across the Domain Workspace***	*- All features of Silver, plus:* *- Programmatic access to all metadata, including data class descriptions, site and source information, and any other caBIG-defined metadata requirements and use information models* ***- Use the caBIG standard or electronic representation of metadata and Common Data Elements***
Information Models	***- No particular information model is used to represent data***	***- Some type of diagrammatic model describing the data relationship is available in electronic format***	***- Information models defined in a standard modeling language such as UML***	*- All features of Silver, plus:* ***- Information models are harmonized with other s across the caBIG Domain Workspace***

Clinical Research IT Infrastructure



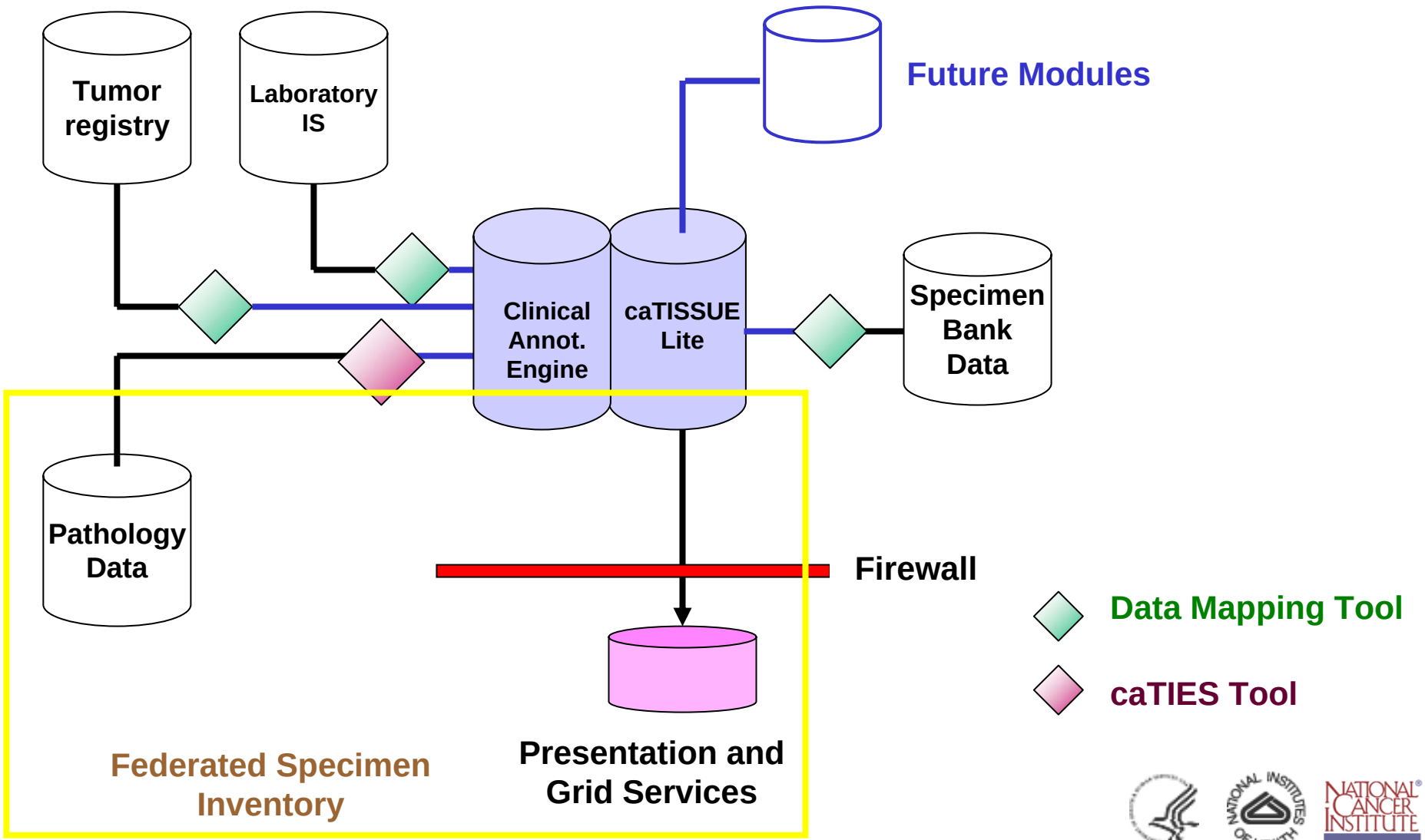
caBIG pilot status - products

Clinical Trials Management Systems

- WWW-based clinical trials data capture and management tools – C3D
  - NCI regulatory reporting
  - Laboratory Interfaces
  - Patient participant registry
  - (caBIG-bronze+ compatibility)
- Financial component model
- Laboratory interface model
- HL7 toolkit (4th quarter 2004)
- E-infrastructure for Investigator Registration (1st quarter 2005)
- Structured protocol model (1st quarter 2005)
- 5 Cancer Centers utilizing C3D infrastructure (2nd quarter 2005)
- C3D clinical data warehouse (2nd quarter 2005)
- E-IND filing (2nd quarter 2005)
- AE reporting module (3rd quarter 2005)

caBIG cancer Biomedical Informatics Grid

Tissue Bank and Pathology Tools Model



caBIG pilot status - products

Tissue Bank and Pathology Tools Workspace

- caTISSUE architecture and use cases
- Federated Tissue Data Set White Paper
- Data Sharing Federation Operational Guidelines (4th quarter 2004)
- caTIES beta release (1st quarter 2005)
- caTISSUE Lite prototype (2nd quarter 2005)
- caTISSUE prototype (2nd quarter 2005)
- External module connector prototype (2nd quarter 2005)
- De-identification reports tool operational (4th quarter 2005)

caBIG pilot status - products

Integrated Cancer Research

- Gene Annotation
  - PIR (2nd quarter 2005)
  - Cancer Molecular Pages (3rd quarter 2005)
  - Function Express (3rd quarter 2005)
  - GoMiner (3rd quarter 2005)
  - HapMap (3rd quarter 2005)
  - SEED (4th quarter 2005)
- Data Analysis and Statistical Tools
  - Distance-Weighted Discrimination (2nd quarter 2005)
  - Magellan (2nd quarter 2005)
  - VISDA (2nd quarter 2005)
  - Gene Pattern (4th quarter 2005)
- Translational (Clinical Integration)
  - TrAPSS (3rd quarter 2005)
- Informatics for Proteomics
  - LIMS (2nd quarter 2005)
  - Q5 (3rd quarter 2005)
  - RProteomics (4th quarter 2005)
- Microarray Repositories
  - caArray (4th quarter 2004)
  - NCI-60 Data Sharing (4th quarter 2004)
  - Zebrafish Mircroarray Data Sharing (4th quarter 2004)
- Pathways
  - Cytoscape/BioPAX/cPath (3rd quarter 2005)
  - QPACA (3rd quarter 2005)
  - Reactome (4th quarter 2005)

Getting involved with caBIG

- Track activities and progress on caBIG Web site at **http://caBIG.nci.nih.gov**
- Attend open meetings to contribute ideas and to better understand caBIG activities.
- Work toward making your applications and solutions caBIG compatible. Current draft guidelines for caBIG compatibility are available on the caBIG Web site
- Download and get familiar with the tools and applications already available on the caBIG Web site.
- Sign up for the caBIG mailing list at on the caBIG Web site to receive important notices about current and upcoming activities.
- Learn more about the National Cancer Institute's existing bioinformatics infrastructure – caCORE (**http://ncicb.nci.nih.gov/core**)